Statistical-based Approach for Indonesian Complex Factoid Question Decomposition

Setio Basuki[1] and Ayu Purwarianti[2]

[1]Informatics Department, Universitas Muhammadiyah Malang, Indonesia
[2]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia
[1]setio_basuki2@yahoo.co.id, [2]ayu@stei.itb.ac.id

Abstract: This research has proposed a method to decompose complex factoid question into several independent questions. The method comprises four stages: (1) classifying input question into several categories such as sub-question, coordination, exemplification, or double question, (2) generating all possible question boundary candidates, (3) selecting the best question boundary, and (4) performing the question decomposition rule using the best question boundary. This study compared several machine learning algorithms in the first stage (complex factoid question classification) and third stage (question decomposition boundary selection). The features used in the classification are specific word lists with its related information including the syntactic features of POS (Part of Speech) tag. For the experiments, we annotated 916 sentences for training data and 226 sentences for testing data. The perplexity of the annotated corpus achieved 1.000586 with 307 Out of Vocabulary (OOV). The complex factoid question classification accuracy reached 93.8% with Random Forest algorithm. The question decomposition boundary selection accuracy achieved 93.80% for sub-question (using Random Forest algorithm), 86.11% for double question (using Random Forest algorithm), 88.23% for coordination (using SMO), and 60.87% for exemplification (using kNN, NB, and RF). A revision rule was provided for the question decomposition boundary selection that improved the accuracy into 97.22% for double question, 94.11% for coordination, and 65.21% for exemplification.

Keywords: complex factoid question, question decomposition, sub-question, coordination, exemplification, and double question.

1. Introduction

Question Answering System (QAS) is Natural Language Processing (NLP) application which is able to provide an answer for user question. Rather than using a search engine application where user has to trace each retrieved document manually in order to find the answer, a QAS provides specific answers and its relevant passage directly. For an open domain QAS, the system usually consists of three subsystems, namely question analysis, passage retrieval, and answer finder. Currently, there are several Indonesian QAS as in [1][9][10][11][14][15] but no research has addressed to the problem of the complex question. Complex question sentence is a sentence that contains more than one independent question. There are several types of complex question sentence such as sub-question, coordination, and exemplification [13]. In providing the answer to the user, the QAS decomposes a complex question into several independent or simple questions and each question is passed to a standard QAS.

Basically, the method to decompose a complex question sentence can be divided into rule-based method and statistical-based method. In the rule-based method, it needs to define manually all possible rules to decompose the complex question [12]. In the statistical-based method, it has to provide the data set for complex question and its decomposition features as a learning resource for a machine learning algorithm. This learning process produces the classifier or rules of complex question decomposition automatically [13]. Both techniques above usually employ lexical, syntactic, or semantic information of the complex question as in

Received: December 15th, 2015. Accepted: June 17th, 2016
DOI: 10.15676/ijeei.2016.8.2.9

[4][5][7][12][13]. Lexical information consists of keyword list including conjunctions and other words or phrases which appear in the complex question. The two main problems in this approach are the number of keywords and ambiguity. In the syntactic-based approach, the POS Tagging Sequence or the Syntactic Parse Tree of the question is commonly used. Semantic-based approach commonly employs the result of semantic analyzer tool such as Predicate Argument Structure (PAS), the semantic representation using First Order Logic (FOL), etc. Currently, there have been some researches related to the English complex question decomposition. Research [4] using Predicate Argument Structure (PAS) and Random Walk Model for Open Domain to decompose open domain complex question, [5] utilizing the PAS to decompose some facts contained in the question, and [7] comparing the decomposition of Open Domain Question using Syntactic-Rule and Semantic approach.

Research [12] decomposed complex English medical question using some rules and keyword list. The rules identified the existence of conjunctions such as "and" and "or" connecting Noun and Noun Phrase in the question and the existence of some triggering words like "such as" and "Including" for detail-required questions. Because the rules were constructed based only on keywords and rules, this study has several drawbacks: (1) not being able to distinguish between conjunctions that can be decomposed and cannot be decomposed, and (2) cannot distinguish between the details that can be decomposed and inseparable details. The accuracy of this study increased in [13] which builds statistical-based decomposition technique using machine learning. The proposed technique used the 3-step Rank-and-Filter: (1) generating all possible question decomposition boundaries, (2) ranking all boundary candidates based on classification probability using machine learning, and (3) performing the decomposition for the highest rank boundary candidate.

The proposed decomposition technique of Indonesian Open Domain complex factoid question was inspired by the results of the study [13]. Our contribution is the lexical and syntactic features of the complex questions both in the complex question classification and boundary selection. Both processes were statistically-modeled using machine learning algorithm. To improve the selection accuracy, some revision rules were also established. Other than the complex question types analyzed in [13] of sub-question, coordination, and exemplification, we also add double question as one of the question types after observing the question data set.

2. Indonesian Complex Question

A. Complex Factoid Question

According to [5][7][13], the complex question is defined as a question asking for several entities, events, or complex relation in a single question. In other words, a complex question contains multiple independent questions where the answer for each question may reside on a different paragraph or even on the different document. In this research, a complex question is defined as a question that contained more than one independent question which forms the structure of sub-question, double question, coordination, and exemplification. The explanation of each Indonesian complex question type is as follow.

- *Sub-question* is identified by conjunction “dan” (and), “atau” (or), “serta” (along), and “/”; that connect 2 independent questions. For example “*Di propinsi manakah terdapat Candi Borobudur dan kapan candi tersebut ditemukan?*” (In what province does Borobudur temple reside and when was that temple built?). The decomposition results for that question are “*Di propinsi manakah terdapat Candi Borobudur?*” (In what province does Borobudur temple reside?) and “*Kapan candi tersebut ditemukan?*” (When was that temple built?).
- *Double Question* is identified by the presence of more than one question word/question expression in the beginning of question. Conjunction for this question includes “dan” (and), “atau” (or), “,” (comma), and “serta” (along). For example “*Pada tanggal berapa dan berapa lamakah Perang Diponegoro berlangsung?*” (On what date and how long did War of Diponegoro occur) contains 2 question expressions. This question can be decomposed into “*Pada tanggal berapa Perang Diponegoro berlangsung?*” (On what date was War of

Diponegoro) and "*Berapa lamakah Perang Diponegoro berlangsung?*" (how long did War of Diponegoro occur).

- *Coordination* is identified by the existence of conjunction that connects some words or phrases in a single question. There are several conjunctions that can be used, namely "dan" (and), "atau", "/", "serta" (along), "dengan" (with), "maupun" (as well as) and "," (comma). In some cases, conjunction "seperti" (such as) also can be used, but it is different from keyword "seperti" (such as) that is used in *Exemplification*. For example "*Kementrian manakah yang bertanggung jawab dalam penyelenggaraan haji dan penetapan awal Ramadhan?*" (Which ministry is responsible for organizing hajj and the deciding the start of Ramadan). The decomposition results are "*Kementrian manakah yang bertanggung jawab dalam Penyelenggaraan Haji?*" (Which ministry is responsible for organizing hajj) and "*Kementrian manakah yang bertanggung jawab dalam penetapan awal ramadhan?*" (Which ministry is responsible for deciding the start of Ramadan).
- *Exemplification* is identified with the presence of an optional phrase in a question. The optional phrase is not tightly related to the main question, it is just for optional explanation. The keywords commonly used '"termasuk" (including), "seperti" (such as), "yaitu" (namely), "contoh" (example), "misalkan" (for example), and the variation of "misalkan" such as "misalnya" and "contohnya". For example "*Apa sebutan untuk hewan yang hanya makan tumbuhan, seperti rumput, daun, dan bunga?*" (What are kinds of animals which only consume plants, such as grass, leaf, and flower?). Decomposition results are "*Apa sebutan untuk hewan yang hanya makan tumbuhan?*" (What are kinds of animals which only consume a plant?), "*Apa sebutan untuk hewan yang hanya makan tumbuhan, seperti rumput?*" (What are kinds of animals which only consume a plant such as grass), "*Apa sebutan untuk hewan yang hanya makan tumbuhan, seperti daun?*" (What are kinds of animals which only consume a plant such as leaf), and "*Apa sebutan untuk hewan yang hanya makan tumbuhan, seperti bunga?*" (What are kinds of animals which only consume plant such as flower).
- *Semantic-dependent* (can not be decomposed) is identified with the presence of semantic-dependent relation among each component. This question cannot be decomposed because the removal of this relation can affect the full meaning of the question. This question commonly contains some conjunctions namely "antara" (between), "menghubungkan" (connect), "sekaligus" (at once) , etc. In this paper, this kind of question is called semantic-dependent.

B. Complex Question Data Source

Complex questions used in this research were obtained from 2 sources. First, the questions were gathered from fourty Indonesian native speakers. Second, we used four English question corpus from previous research, two universities, and TREC. Not all the question provided in these two sources are complex factoid question. Thus, we separated the complex factoid questions with other question types. For every English complex question, it must be translated to Indonesian language, adjust its syntax, and group the question according to its type. The complex question dataset is shown in Table 1.

Table 1. Complex Question Data Source

Data Sources	Selected Question Dataset	Complex Factoid Question
Text Retrieval Conference (TREC)	1829	58
Carnegie Mellon University (CMU)	1568	89
University of Illinois	5000	573
Indonesian Native Speakers (40 Students)	800	737
Simple or Complex Question Classification by [2]	4542	309
Total	13659	1689
Total (After Revision)	-	1142

In this research, two question corpus were built. The first corpus is used for the complex question classification. It consists of question sentence and its label. Each complex question in this corpus was labeled with question type such as simple question, sub-question, double question, coordination, exemplification, and semantic-dependent. The example of each question type is shown in Table 2.

Table 2. Example of Question and Its Type in Complex Question Classification Corpus

No.	Question Type	Question Example
1	Simple Question	*Berapa persen peningkatan jumlah bencana alam yang terjadi sejak tahun 2010-2013?:simple* *(How many percent was the increasing number of natural disasters that have occurred since the year 2010-2013?: Simple)*
2	Double Question	*Kapan pertama kali, untuk siapa, dan berapa kali Ir. Soekarno menulis surat selama di asingkan di Ende?:double* *(When was the first time, to whom, and how many times did Ir. Soekarno write while being in exile in Ende?:double)*
3	Coordination	*Siapakah seniman yang sangat dipengaruhi oleh arsitektur dan budaya Montevideo?:coordination* *(Who is the artist who is greatly influenced by the architecture and culture of Montevideo?:coordination)*
4	Exemplification	*Disebut apakah suatu danau yang terbentuk dari tertahanya air oleh bahan lepas, seperti runtuhan gunung, moraire ujung gletser, dan aliran lava?:exemplification* *(What is the name of the lake which is formed from water retention by loose material, such as the ruins of the mountain, moraine ends of glaciers, and flows of lava?:exemplification)*
5	Sub-Question	*Apa ponsel pertama yang mempunyai fitur kamera dan perusahaan apa yang membuatnya:?sub-question* *(What is the first cell phone which has camera feature and what company produced it?:sub-question)*
6	Semantic-Dependent	*Apa nama acara komedi yang terdiri dari seorang Artis bernama Nora Desmond, Sekretaris bernama Wiggins, dan seorang ibu rumah tangga bernama eunice?:semantic-dependent* *(What is the name of comedy show consisting of artists named Nora Desmond, the Secretary named Wiggins, and a housewife named Eunice?: Semantic-dependent)*

Table 3. Example of Question and Its Boundary in Question Decomposition Boundary Corpus

No.	Annotation Class	Annotation Example
1	Double Question Positive	*[Kapan pertama kali, untuk siapa, dan berapa kali] Ir. Soekarno menulis surat selama di asingkan di Ende?* *([When was the first time, to whom, and how many times] did Ir. Soekarno write while being in exile in Ende)*
	Double Question Negative	*[Kapan dan di] negara mana pertama kali muncul wabah virus Ebola* *([When and in] what country the Ebola virus came up)*
2	Coordination Positive	*Siapakah seniman yang sangat dipengaruhi oleh [arsitektur dan budaya] Montevideo?* *(Who is the artist who is greatly influenced by the [architecture and culture] of Montevideo.)*
	Coordination Negative	*Disebut apa proses gabungan antara [evaporasi dan transpirasi]* *(What process is produced from a combination between [evaporation and transpiration])*
3	Exemplification Positive	*Disebut apakah suatu danau yang terbentuk dari tertahanya air oleh bahan lepas, seperti [runtuhan gunung, moraire ujung gletser, dan aliran lava]?* *(What is the name of the lake which is formed from water retention by loose material, such as [the ruins of the mountain, moraine ends of glaciers, and lava flows])*
	Exemplification Negative	*Apa nama perusahaan minyak Amerika Serikat yang sekarang terpecah menjadi beberapa perusahaan seperti [Exxon Mobil, Chevron, dan Amoco]?* *(What is the name of the US oil company that is now split into several companies such as [Exxon Mobil, Chevron, and Amoco])*

The second corpus is used for determining the question decomposition boundary. This corpus contains boundary "[" and "]" on the complex question as a decomposition reference for the 3 types of complex question, namely Double Question, Coordination, and Exemplification. For each question type, we provided positive and negative annotation sample. Positive annotation is an example of the correct boundary, whereas a negative annotation is an example of the wrong boundary. The entry sample of question decomposition boundary corpus is shown in Table 3.

3. Complex Question Decomposition

Decomposition process is started when the system receives complex question from user. The question is classified into several complex question types. If the question is classified as an independent (simple question) or sub-question, then the question will not be forwarded to the decomposition module. Independent question can be forwarded directly to baseline QAS and sub-question can be directly decomposed. If the question is classified as a double question, coordination, or exemplification, then this question will be forwarded to the decomposition module which consists of three phases namely decomposition boundary generation, decomposition boundary selection, and performing decomposition rule based on the boundary. The entire process is illustrated in Figure 1 below.

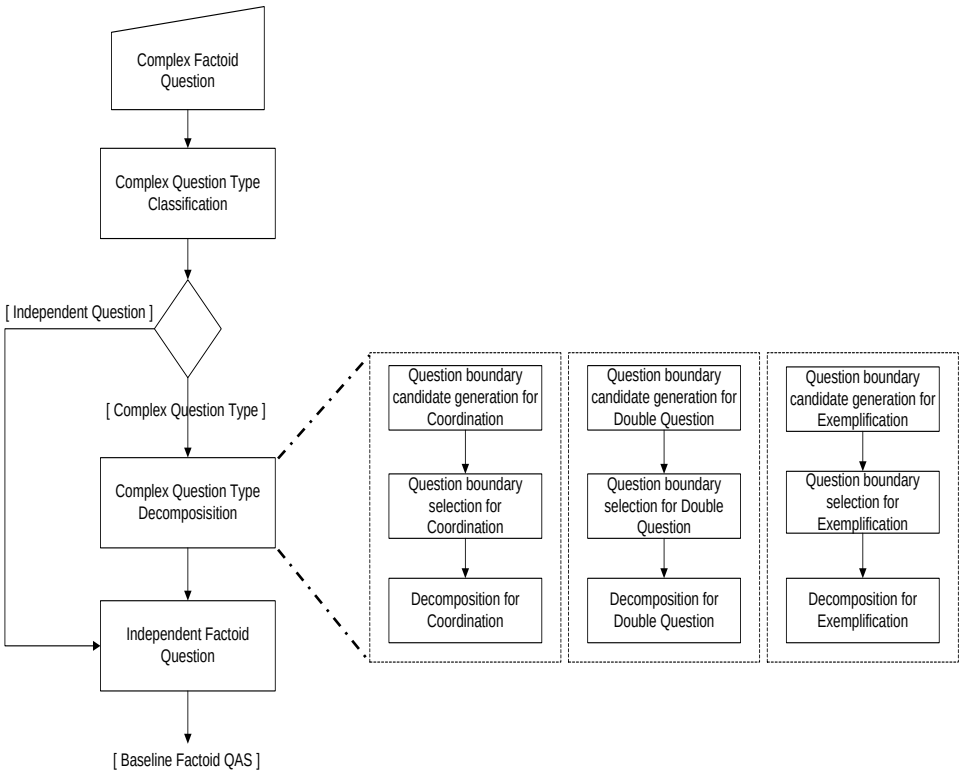

Figure 1. Complex Question Decomposition System Architecture

A. Complex Question Classification

Complex question classification is done using several machine learning algorithms such as Sequential Minimal Optimization (SMO), Naive Bayes (NB), k-Nearest Neighbor (kNN), C4.5, and Random Forest (RF). Features for the classification consist of several word lists and other information related with the word in the word list. There are three word lists employed, namely the conjunction word list, the interrogative word list, semantic-dependent word list and exemplification word list. We also used a special punctuation of “comma” which is usually

exists in a complex question sentence. The complete features used for the complex question classification and the examples for sentence “Apa pegunungan yang terletak antara Arkansas dan sungai Missouri?” (What mountains are among Arkansas and river of Missouri?) are shown in Table 4.

Table 4. Features for Complex Question Classification

No	Feature	Details
Apa pegunungan yang terletak antara arkansas dan sungai missouri? *(What mountains are among Arkansas and river of Missouri?)*		
1	Conjunction’s type	Conjunction in the complex question *Feature: “and”*
2	Conjunction’s index	Conjunction position relatively early in the question indicates the double question. The position is calculated from the mostly left token. *Features: "no", because there is conjunction within the first 5 words*
3	POSTag of the word (not a comma) exactly before and after the conjunction	Coordination is generally surrounded by words that have the same Part of Speech Tag (POSTag) on the left and right conjunctions *Fitur: “NNP” (Arkansas) dan “NN” (river)*
4	The existence of interrogative words on the left and right of conjunction	Check 3 words on the left and right conjunction. *Features: "no", there is no interrogative word within 3 words to the left and right conjunctions*
5	The presence of semantic-dependent word	The semantic-dependent word indicates that complex question does not require decomposition *Features: "yes", the question contains the keyword "among"*
6	Distance between conjunction index and semantic-dependent index	The presence of semantic-dependent word before conjunctions indicates that the question does not require decomposition *Feature: “2", distance between conjunction with the word "among"*
7	The existence of Exemplification word	The presence of this word indicates that the question is Exemplification *Features: "no", there is no an Exemplification word*
8	The existence of a comma before the exemplification word	The presence of this word indicates that the question is Exemplification *Features: "no", there is no Exemplification word and no comma in the sentence*

B. Question Boundary Candidate Generation

This module is responsible to generate all candidates of question boundary by establishing all possible boundaries for each complex question. Candidates are all possible combinations of the boundary mark "[" as the beginning and "]" as the end attached to the question. One complex question could have many candidates according to the number of tokens (word or punctuation) it has. This process is built using the rule that each boundary candidate used as test data in the selection process. Examples of question boundary candidate generation result are shown in Table V which shows the most appropriate boundary (bold-typed ones) as a decomposition reference.

Table 5. Examples of Question Boundary Candidate Generation Results

No.	Question Type	Generated Candidates
1	Double Question	**Question:** *Kapan, dimana, dan untuk berapa lama Perang Salib pertama kali terjadi?* *(When, where, and for how long did the first Crusade occur?)* **Classification Result:** Double Question **Question Boundary Candidates:** *Candidate no 1. [Kapan , dimana , dan untuk berapa lama Perang Salib pertama kali terjadi]* *Candidate no 2. Kapan [, dimana , dan untuk berapa lama Perang Salib pertama kali terjadi]* *...* ***Candidate no 21. [Kapan , dimana , dan untuk berapa lama] Perang Salib pertama kali terjadi*** *...* *Candidate no 32. Kapan , dimana [, dan untuk] berapa lama Perang Salib pertama kali terjadi*
2	Coordination	**Question:** *Siapakah nama gubernur pertama provinsi Jakarta , Jawa barat, dan Jawa Tengah?* *(What is the name of the first Governor of Jakarta, West Java, and Central Java?)* **Classification Result:** Coordination **Question Boundary Candidates:** *Candidate no 1. [Siapakah nama gubernur pertama provinsi Jakarta , Jawa barat , dan Jawa Tengah]* *Candidate no 2. Siapakah [nama gubernur pertama provinsi Jakarta , Jawa barat , dan Jawa Tengah]* *...* *Candidate no* ***6. Siapakah nama gubernur pertama provinsi [Jakarta , Jawa barat , dan Jawa Tengah]*** *...* *Candidate no 20. Siapakah nama gubernur pertama provinsi Jakarta , Jawa barat [, dan Jawa] Tengah*
3	Exemplification	**Question:** *Disebut apakah pengembangan beragam kompetensi peserta didik tentang seni, termasuk pengetahuan, pemahaman, analisis, evaluasi, apresiasi, dan kreasi?* *(What do we call the action to develop the variety of students' competency about art, including knowledge, understanding, analysis, evaluation, appreciation, and creation?)* This complex question can be split first exactly at Exemplification word "termasuk" (including), to simplify the boundary selection process. **Classification Result:** Exemplification **Question Boundary Candidates:** *Candidate no 1. [seni , termasuk pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]* *Candidate no 2. seni [, termasuk pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]* *Candidate no 3. seni , [termasuk pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]* *Candidate no* ***4. seni , termasuk [pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]*** *...* *Candidate no 13. seni , termasuk pengetahuan , pemahaman , analisis , evaluasi , apresiasi [, dan kreasi]*

C. *Question Boundary Selection*

Selection is done by building a binary classification: positive and negative using machine learning algorithms. The classification process is not to determine the membership of a candidate to get into a class, but rather determining the probability value of each candidate to be categorized into positive class. The highest probability candidate to be categorized as a positive class would be selected as a decomposition reference. Therefore, in this step, positive and negative annotated training data are provided for double-question, coordination and exemplification questions. Sub-question type do not run into this process and it would be directly forwarded to the decomposition module. Table VI shows candidate examples and its probability value of having positive class. The bold-typed candidate is the candidate which has the highest probability value.

Table 6. Example of Question Boundary Candidate and Its Probability Score

No.	Question Type	Question Candidates
1	Double Question	*Candidate no 1. [Kapan , dimana , dan untuk berapa lama Perang Salib pertama kali terjadi]* Classification Result: Probability to be positive class : 0.6086, Probability to be negative class : 0.3913 ... ***Candidate no 21. [Kapan , dimana , dan untuk berapa lama] Perang Salib pertama kali terjadi*** Classification Result: Probability to be positive class : 0.9461, Probability to be negative class : 0.0538 ... *Candidate no 32. Kapan , dimana [, dan untuk] berapa lama Perang Salib pertama kali terjadi* Classification Result: Probability to be positive class : 0.0352, Probability to be negative class : 0.9647
2	Coordination	*Candidate no 1. [Siapakah nama gubernur pertama provinsi Jakarta , Jawa barat , dan Jawa Tengah]* Classification Result: Probability to be positive class : 0.0, Probability to be negative class : 1.0 ... ***Candidate no 6. Siapakah nama gubernur pertama provinsi [Jakarta , Jawa barat , dan Jawa Tengah]*** Classification Result: Probability to be positive class : 0.9112, Probability to be negative class : 0.0887 ... *Candidate no 20. Siapakah nama gubernur pertama provinsi Jakarta , Jawa barat [, dan Jawa] Tengah* Classification Result: Probability to be positive class : 0.0030, Probability to be negative class : 0.9969
3	Exemplification	*Candidate no 1. [seni , termasuk pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]* Classification Result: Probability to be positive class : 0,0055, Probability to be negative class : 0,9945 … ***Candidate no 4. seni , termasuk [pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]*** Classification Result: Probability to be positive class : 0,9852, Probability to be negative class : 0,0147 ... *Candidate no 13. seni , termasuk pengetahuan , pemahaman , analisis , evaluasi , apresiasi [, dan kreasi]* Classification Result: Probability to be positive class : 0,0165, Probability to be negative class : 0,9834

Table 7. Features for Question Boundary Selection on Double-Question Type

No	Feature	Details
Example for features 1, 2, 3, 4, and 5: *[Kapan pertama kali dan berapa kali] Soekarno menulis surat selama di Eden?* *([When was the first time and how many times] Soekarno wrote a letter during in Eden?)*		
1	Conjunction's type	Conjunction in the complex question *Feature: "and"*
2	The existence of interrogative words on the left and right conjunction	The double question has a question word or a question expression which resides on the left and right of conjunctions or a comma at the beginning of the question. *Features: "yes" and "yes", there are interrogative words on the left and right conjunction.*
3	Part of Speech Tag (POSTag) of a word before the right boundary	Double Question can have the question words, dates, or adjective just before the right boundary *Feature: "NN" for word "times" which represents dates*
4	Part of Speech Tag (POSTag) of the word after the left boundary	The word just after the left boundary which is generally a question word or question expression *Fitur: "WP" for word "When"*
5	Part of Speech Tag (POSTag) of the word after the right boundary	The word after the right boundary generally is not the dates, adjectives, or question words Example: *Feature: "NNP" for word "Soekarno"*
Example for features 6, 7, 8, 9, and 10: *Untuk berapa lama, kapan [, dan oleh siapa] Soekarno diasingkan di Bengkulu?* *(For how long, when [, and by whom] Soekarno was exiled in Bengkulu?)*		
6	The existence of interrogative words before the left boundary	Not to be the best candidate if there is a question word before the left boundary. The question words should be part of the best candidate Example: *Features: "yes", before the left boundary there is 2 question word "when" and "how"*
7	The existence of a comma before the left boundary	Not to be the best candidate if there is a comma before the left boundary because a comma indicates that there are components that should be mentioned Example: *Features: "yes", because there is a comma before the left boundary*
8	The presence of a comma after the left boundary	The existence of comma just after the left boundary indicates that the candidate is not the best decomposition reference Example: *Features: "yes" because there is a comma after the left boundary*
9	The existence of a comma just before the conjunction	The existence of a comma just before the conjunction indicates that the candidate is not the best decomposition reference because it does not contain detail components *Features: "yes" because of the coma right before conjunctions*
10	The presence of one or more words before left boundary	Not to be the best candidate if there is a word before the left boundary *Features: "yes" because there are several words before the left boundary*
11	The existence of question word just before the right boundary	Double Question can have a question word just before the right boundary as the rightmost limit decomposition reference *Features: "yes", there is a question word "whom"*
Example for feature 12: *[Dimana dan berapa lama] Soekarno diasingkan Belanda untuk pertama kalinya?* *([Where and for how long] was Sukarno exiled by Netherlands for the first time?)*		
12	The existence of certain adjective before the right boundary	Double Question can have a certain adjective just before the right boundary Example: *Features: "yes", there is the adjective "long" right before the right boundary*
Example for feature 13: *[Kapan, oleh siapa, dan berapa hari] Perjanjian Linggarjati dilaksanakan?* *([When, by whom, and how many days] was Linggarjati Agreement organized?)*		
13	The existence of a word which represents a date before the right boundary	Double Question can have a word that represents a date just before the right boundary *Features: "yes" includes the word "days" just before the right boundary*

The feature used in the question boundary selection for each complex question type is similar. It consists of word list feature of surrounding words related with the boundary. There are several differences among the three models of double question, combination and exemplification based on the question pattern that we observed. For the double question, we proposed the complete features as shown in Table 7.

Table 8. Features for Question Boundary Selection on Coordination Question Type

No	Feature	Details
Example for features 1, 2, 3, 4, and 5: *Pada abad berapakah kerajaan [Tarumanagara dan Kutai] menguasai Nusantara?* *(In what century did the empire [Tarumanegara and Kutai] rule the Nusantara?)*		
1	Conjunction's type	Conjunction in the complex question *Feature: "and"*
2	Part of Speech Tag (POSTag) of the word before the conjunction	Conjunction in coordination generally connects two words with the same POSTag *Feature: "NNP" for Tarumanegara*
3	Part of Speech Tag (POSTag) of the word after the conjunctions	Conjunction in coordination generally connects two words with the same POSTag *Feature: "NNP" for Kutai*
4	Part of Speech Tag (POSTag) of the word just before the left boundary	Determine POSTag of the word just before the left boundary of valid candidate Example: *Feature: NN for "kerajaan"*
5	Part of Speech Tag (POSTag) of word just after the right boundary	Determine POSTag of the word just after the right boundary of valid candidate Example: *Feature: VBT for "rule"*
Example for features 6 and 7: *Apa nama Laut yang terletak di antara [Israel, Daerah Otoritas Palestina, dan Yordania]?* *(What is the name Sea that lies between [Israel, the Palestinian Authority Region, and Jordan]?)*		
6	The presence of a comma before the left boundary	The presence of a comma before the left boundary indicates that this candidate cannot be used as decomposition reference *Feature: "no" because there is no comma before the left boundary*
7	The presence of a semantic-dependent word in the question	Not to be the best candidate if it contains semantic-dependent word *Features: "yes", because there is a semantic-dependent keyword "between"*
Example for feature 8 and 9: *Pulau apa [yang menjadi pembatas antara Indonesia dan Papua Nugini]?* *(What island [which becomes the boundary between Indonesia and Papua New Guinea]?)*		
8	The presence of a question word after the left boundary	The valid Coordination candidate does not contain a question word after the left boundary. *Feature: "no"*
9	The existence of a question word before the left boundary	The valid Coordination candidate does not contain a question word just before the left boundary *Feature: "yes", there is a question word "What"*
Example for feature 10 and 11: *Apa nama sungai yang melalui Colorado, [Kansas, dan Oklahoma]?* *(What is the river which flows through Colorado, [Kansas, and Oklahoma]?)*		
10	The existence of a comma before the left boundary	The valid Coordination candidate does not contain a comma before the left boundary, because there are still some details that have not been accommodated *Feature: "yes"*
11	The existence of a comma after the left boundary	The valid Coordination candidate does not contain a comma just after the left boundary, because there are still some details that have not been accommodated *Feature: "no"*
Example for feature 12 and 13: *Di mana pusat gempa Aceh khususnya garis [Lintang dan Bujur] pada tahun 2004?* *(Where is the center of the earthquake in Aceh, especially the line [Latitude and Longitude] that occurred in 2004?)*		
12	Part of Speech Tag (POSTag) of the word after the left boundary	Determine POSTag of the word after the left boundary of valid candidate *Feature: "NN" for "Latitude"*
13	POS of the word before right	Determine POSTag of the word before the right boundary of valid

No	Feature	Details
	boundary	candidate *Feature: "NN" for "Longitude"*
Example feature 14 and 15: *Siapa novelis mata-mata yang menjabat sebagai koresponden kantor berita [Reuter dan Times] di London?* *(Who is the spy novelist who served as a corresdence for the news agency [Reuter and Times] in London?)*		
14	The presence of preposition after the right boundary	The valid Coordination candidate may contain the prepositions just after the right boundary *Feature: "yes" for preposition "in"*
15	The existence of a question words after the right boundary	If the candidate contains a question word after the right boundary, then it is not a valid Coordination candidate. *Feature: "no", there is no question word after the right boundary*
Example for feature 16: *Di jalan apa di Tokyo yang gemerlap dipenuhi dengan [department store dan klub malam]?* *(In what glittered street in Tokyo which is filled with the [department stores and night clubs]?)*		
16	The existence of certain phrase after the conjunction	It is to identify that the details are not a single word but a phrase *Features: phrase "department store" and "nightclub"*

For coordination question type, some features are inspired by previous research [13], particularly numbers 1, 2, 3, and 7. The remaining features are new which we propose according to the character of the Indonesian coordination question. The complete features are shown in Table 8.

For exemplification question type, some features for the selection are also inspired by previous research [13], particularly numbers 1 and 12. The remaining features are new which we propose according to the character of the Indonesian Exemplification question.

Table 9. Features for Question Boundary Selection on Exemplification Question Type

No	Feature	Details
Example for feature 1 and 2: *Disebut apa suatu norma yang diikuti hanya berdasar adat kebiasaan masyarakat, misalnya [cara mengangkat topi, cara duduk, dan cara makan]?* *(What is the norm which is formed from community's habits and traditions, such as [how to lift the cap, how to sit, and how to eat]?)*		
1	The presence of an Exemplification word before the left boundary	The valid candidate has an Exemplification word before the left boundary Feature: "yes", there is exemplification word "such as" in the question.
2	The availability of a comma just after the left boundary	The valid candidate does not have a comma after the left boundary Feature: "no", there is no comma just after the left boundary
3	The presence of a comma just before the left boundary	The valid candidate does not have a comma just before the left boundary Feature: "no", there is no comma just before the left boundary
Example for feature 4: *Disebut apa data yang mempunyai nilai berupa pecahan, misalnya pengukuran panjang, luas, isi, waktu [, dan berat]?* *(What type of data which has fraction value, for example, length measurements, area, content, time [, and weight]?)*		
4	The presence of a comma just after the right boundary and then it is followed by a conjunction	Not to be the valid candidate if there is a comma just after the left boundary and the presence a conjunction after it *Features: "yes" because there is a comma just after left boundary then immediately followed by the conjunction.*
Example for feature 5: *Apa nama metode pengendalian hama, seperti [serangga dan jamur]?* *(What is the name of pest control methods, such as [insects and fungi])*		
5	Part of Speech Tag (POSTag) of a words before the conjunction	Determine POSTag of the word before the conjunction of valid candidates *Features: "NN" for "insect"*
Example for feature 6: *Disebut apa mekanisme transpor bahan bahan mineral di dalam tumbuhan , [misalnya air mineral dan hasil fotosintesis]?* *(What is the mineral transport mechanism inside of plants, [for example, mineral water and photosynthesis result]?)*		
6	The existence of an exemplification word after the left boundary	The valid candidate does not contain an Exemplification word before the left boundary *Features: "yes", there is Exemplification keywords after the left boundary*
Example:		

No	Feature	Details
Disebut apa peta yang menggambarkan kevariasian jenis data tanpa memperhitungkan jumlahnya , contohnya [peta tanah , peta budaya , dan peta agama] *What type of map which illustrates data variations regardless of the amount, for example [soil maps, cultural maps, and religious map]* *Features: "yes" because the existence of the comma "," right before left boundary*		
7	The existence of a comma just before the left boundary	The valid candidate does not contain a comma just before the left boundary *Feature: "no"*
8	The presence of a comma just after the left boundary	The valid candidate does not contain a comma just after the left boundary *Feature: "no"*
Example for feature 9 and 10: *Disebut apa bidang yang mengajarkan orang untuk berinveatasi, misalnya [penanaman modal atau saham] di perusahaan?* *(What is the field that teaches people to invest, such as [investment or stock] in the company?)*		
9	The presence of a comma before the Exemplification word	The valid candidate has a comma before the Exemplification word *Features: "yes" , there is a comma just before the Exemplification word*
10	The presence of a preposition after the right boundary	A valid Candidate may contain a preposition after the right boundary Example: *Features: "yes", there is a preposition after right boundary*
Example for feature 11 and 12: *Disebut apa bukti sejarah yang merupakan hasil garapan tangan manusia, seperti [Candi, Patungs, dan Perkakas]* *(What was historical evidence made by human hands called, such as [Temple, Sculpture, and Tools])*		
11	Part of Speech (POSTag) of the word before the right boundary	Determine POSTag of the word before the right boundary of valid candidate *Features: "NN" to the word "Tools"*
12	Part of Speech Tag (POSTag) of the word after the right boundary	Determine POSTag of the word after the right boundary of valid candidate *Features: "Null" because there are no word after the right boundary*

D. *Performing Decomposition Rule*

Table 10. Performing Decomposition Rule on Complex Question

Double Question
Best Candidate:
Candidate Number 21:
[Kapan , dimana , dan untuk berapa lama] Perang Salib pertama kali terjadi
([When, where, and for how long]did the first Crusade occur?)
Decomposition Result:
Kapan Perang Salib pertama kali terjadi (When did the first Crusade occur)
Dimana Perang Salib pertama kali terjadi (Where did the first Crusade occur)
Untuk berapa lama Perang Salib pertama kali terjadi (For how long did the first Crusade occur)
Coordination
Best Candidate:
Candidate Number 6:
Siapakah nama gubernur pertama provinsi [Jakarta , Jawa barat , dan Jawa Tengah]
(What is the name of the first Governor of [Jakarta, West Java, and Central Java])
Decomposition Result:
Siapakah nama gubernur pertama provinsi Jakarta (What is the name of the first Governor of Jakarta)
Siapakah nama gubernur pertama provinsi Jawa barat (What is the name of the first Governor of West Java)
Siapakah nama gubernur pertama provinsi Jawa Tengah (What is the name of the first Governor of Central Java)
Exemplification
Best Candidate:
Candidate Number 4:
termasuk [pengetahuan , pemahaman , analisis , evaluasi , apresiasi , dan kreasi]?
(including [knowledge, understanding, analysis, evaluation, appreciation, and creation])
Decomposition Result:
seni , termasuk pengetahuan? (art, including knowledge)
seni , termasuk pemahaman? (art, including understanding)
seni , termasuk analisis? (art, including analysis)
seni , termasuk evaluasi? (art, including evaluation)
seni , termasuk apresiasi? (art, including appreciation)
seni , termasuk kreasi? (art, including creation)

After the question boundary is defined, the decomposition rules are performed to decompose the complex question into more than one independent question. All words or phrases in the boundary separated by conjunctions (including comma) are decomposed, then each would be combined with the rest of the question components. For double question and coordination types, the number of the decomposed independent question is equal to the number of words or phrases in the boundary separated by conjunctions or commas. For exemplification type, the number of independent question that could be formed is the same as the previous question types plus 1, because there have been independent main questions. Table 10 is an example of the decomposition process which is a continuation of the process of candidate generation in Table 5 and candidate selection in Table 6.

4. Experiment

The evaluation was performed on each component which has been mentioned on decomposition system architecture above. The bi-gram language model perplexity value was evaluated to ensure whether the built corpus has a good language model variety distribution or not. The performance of the system was also evaluated to predict the class of complex question that was given by the user. Decomposition performance evaluation was measured by the accuracy of the system in selecting the best question boundary candidate for decomposition reference in the selection process.

A. Language Model and Perplexity

This section shows the evaluation results of the built corpus according to bi-gram language model. Because the corpus was built from complex question collection, the evaluation was performed by comparing perplexity value and Out of Vocabulary (OOV) of the corpus with and without using the question mark "?".

Table 11. Corpus Bigram Perplexity

Number of Question	Number of Words	Perplexity (with "?")	Perplexity (without "?")	OOV (with "?")	OOV (without "?")
1142	10174	1,000613	1,000586	314	307

The experiment showed that perplexity value and OOV involve question mark bigger than the one without using question mark.

B. Experiments on Indonesia Complex Question Type Classification

In this experiment, 931 training data and 226 testing data were used. Table XII shows the distribution of each complex question type. According to Figure 2, the model built by Random Forest algorithm provided the highest accuracy, 93.8%; while the lowest value was obtained by using Naive Bayes algorithm.

Table 12. Training and Testing Data

No.	Complex Question Type	# Training Data	#Testing Data
1.	Sub Question	167	226
2.	Double Question	103	
3.	Coordination	178	
4.	Semantic-Dependent Coordination	151	
5.	Exemplification	100	
6.	Simple Question	233	
	Total	916	

In the model built by the Random Forest, there were 14 testing data having error classification result. It mostly occurred in 10 coordination questions which were classified as semantic-dependent, and vice versa. This error was caused by the inability of the keyword list as one of the classification features to distinguish the meaning of conjunctive relationship of some words or phrases. For example, *"Apa nama program televisi yang dipandu oleh Gading*

Martin dan Andhika Pratama?" (What is the name of the television program which is hosted by Gading Martin and Andhika Pratama?). This question can have more than one meaning: 1) "What is the name of the television program hosted by Gading Martin" and "What is the name of the television program hosted by Andhika Pratama", or 2) the television program is hosted by both ("Gading Martin and Andhika Pratama").

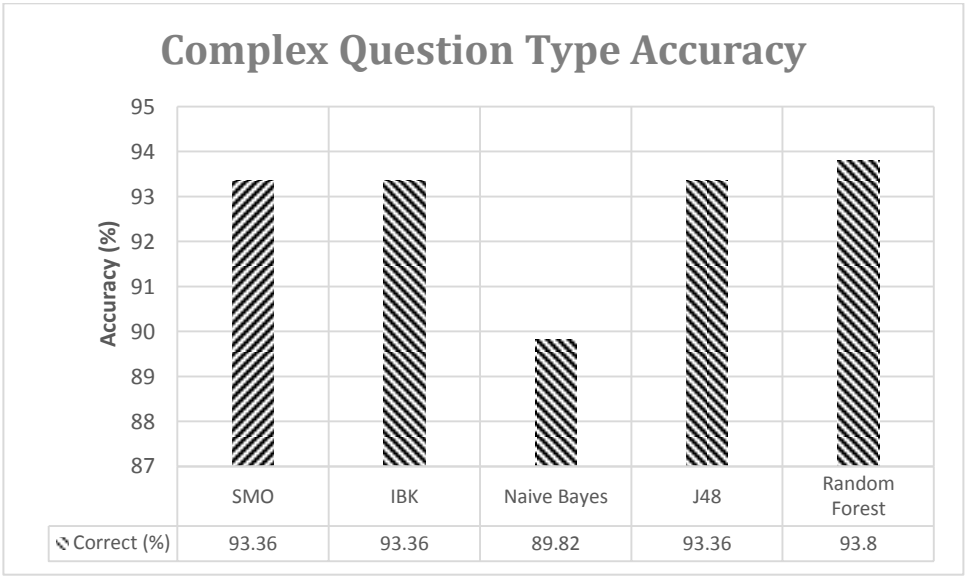

Figure 2. Accuracy Comparison for Complex Question Type Classification

The rest misclassification occurred in 4 questions with 1 exemplification question and 2 sub-questions were classified as semantic-dependent. Also, there was 1 double question classified as sub-question. This error was caused by training data variation that led to the increase in the similarity between the one questions type and others.

C. Experiments on Question Boundary Selection

This experiment aimed to evaluate the accuracy of each model, in selecting the best candidate of question boundary as decomposition reference. There were 5 machine learning algorithms employed such as SMO, Ibk, Naive Bayes, J48, and Random Forest. Each machine learning algorithm was used to build the model for 3 question types of double question, coordination, and exemplification. The experimental result is shown in Table 13 below.

Table 13. Question Boundary Selection Accuracy

	SMO	Ibk	Naive Bayes	J48	Random Forest
Double Question	83.33	80.56	72.22	83.33	86.11
Coordination	88.23	58.82	47.06	73.53	70.59
Exemplification	43.48	60.87	60.87	43.48	60.87

Table 13 shows that the highest accuracy to determine the question boundary on double question was achieved by Random Forest and the question boundary on coordination were achieved by SMO. For the Exemplification type, IBk, Naive Bayes, and Random Forest produced the same accuracy. Set of rules were also added to revise the results since there was an error pattern on the result. The rule for the double question was used to detect whether there was a word which represents time expression or an adjective after the right boundary. In this case, particularly the adjective means a word that represents tall, small, far, long, etc. The rule

for coordination was employed to identify whether the left or the right of the boundary was still a noun. This rule was also applied to the exemplification because it has the same characteristics.

Error in selecting the question boundary on double question can be divided into several types as the followings: (1) the candidate with an adjective before the right boundary such as "berapa tinggi (how tall)" and (2) the candidate with the word representing a date before the right boundary such as "berapa bulan (how many month)". As for the coordination, the incorrect classification was on the candidates which connect a phrase with a word (non-linear list) or one phrase with another phrase such as (1) “Tari Piring, Tari Payung, dan Tari Lilin (Piring Dance, Payung Dance, and Lilin Dance)”, (2) “Kawah Putih dan Tangkuban Perahu”, (3) “Jakarta, Jawa barat, and Jawa Tengah (Jakarta, West Java, and Central Java)”, (4) “Indonesia Raya dan Indonesia Pusaka (Indonesia Raya and Indonesia Pusaka)”. For exemplification, the incorrect classification was on the candidates which connect the collection of words or phrases that are not linear like “infeksi pada oviduk atau infeksi uterus (infection which reside in Oviduct and Uterine infection)”, “setahun yang lalu, setengah tahun yang lalu, seminggu yang lalu, dan hari ini (one year ago, half years ago, one week ago, and today)”, and “bagaimana orang menanamkan modal dan membeli rumah di perumahan (how people invest and buy houses in the housing complex)”. In addition, the incorrect classification was also on candidates with a word or phrase which acts as the quantifier for each detail such as "lembaga perkawinan dan agama (the institution of marriage and religion)".

Certain features have basically been provided at the training data to be able to accommodate the cases mentioned above. Because the classification was done not to determine the candidates membership but to determine the highest classification probability, the inadequacy of training sample has made the candidates who should be chosen to be the decomposition reference not get the highest scores. For that reason, the revision rule was established to fix the question boundary selection result. Rules that have been built were used to improve the accuracy of selecting the question boundary. Therefore, the boundary selection process was first done using machine learning, then the rule would check the position of the boundary before decomposition was performed on the best candidate. For double question type, these rules detect whether after the right boundary there still appeared a word related to the time (seconds, minutes, hours, day, week, month, quarter, season, year, etc) or an adjective (high, much, much, long, long, short, etc). The adjective here generally is an adjective that could be measured. For coordination and exemplification types, the rule is used to identify a noun or noun phrase before or after the boundary.

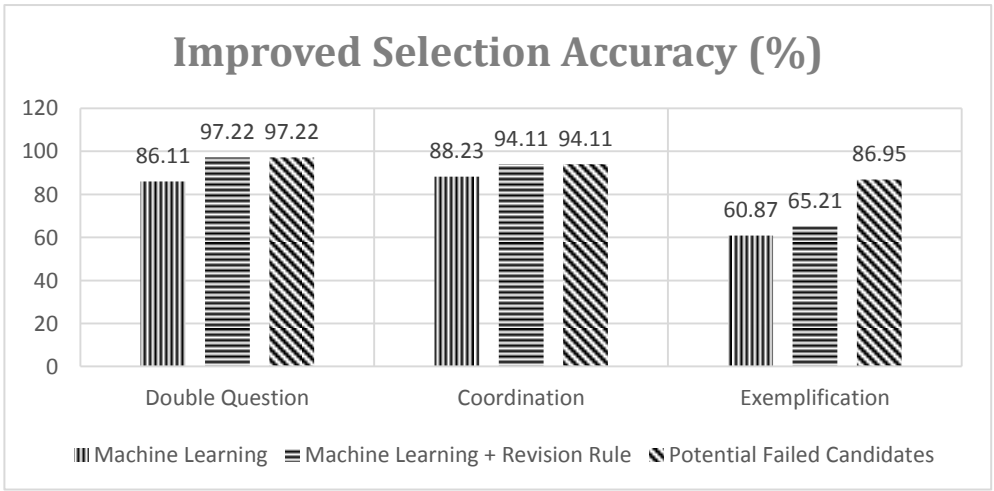

Figure 3. Accuracy Comparison for Question Boundary Selection using Machine Learning and Hybrid (Machine Learning combined with Revision Rule)

Figure 3 indicated an increase of accuracy after providing revision rule for all types of question. For the exemplification, there was an increase about 5% from 60.87% to 65.21%. However, adding the candidate that potentially got the correct answer from the baseline QAS (potential-failed candidate), the accuracy reached 86.95%.

D. Relaxed Match

A relaxed match evaluation was also conducted to obtain the classification probability value of the question boundary should be chosen by the system. This candidate was annotated manually and it was part of whole candidates that can be generated by a single complex question. The purpose of this evaluation was to make sure that the best candidate had high probability value to be the decomposition reference. Therefore, our proposed selection features were considered representative to be used in decomposition system.

This experiment showed that almost all best candidates that should be chosen as the decomposition reference resided on the top three according to its classification confidence. This proved that the classification features are representative for the selection. However, in some test data, the best candidate was not chosen by the system because of lacking samples in the training data that represent candidate's structure.

Table 14. Relaxed Match Evaluation

Number of the best candidate probability that should be chosen as decomposition reference	SMO	IBk	Naïve Bayes (NB)	J48	Random Forest (RF)
Double question, total testing data: 36					
Rank 1 to 3	35	35	35	35	35
Rank > 3	1	1	1	1	1
Coordination, total testing data: 34					
Rank 1 to 3	32	32	32	32	32
Rank > 3	2	2	2	2	2
Exemplification, total testing data: 34					
Rank 1 to 3	32	32	32	32	32
Rank > 3	2	2	2	2	2

5. Conclusion

In this study, the decomposition system for Indonesian complex factoid question has been developed to accommodate the complex question which contains two or more independent questions. 1142 questions for all complex question types which were obtained from previous relevant research and volunteers were used. The core of the decomposition method used in this research was utilizing the probability or classification confidence for every possible question boundary which was generated from a single complex question. The candidate with the highest probability value was used as decomposition reference. Because the best candidate was selected based on the probability value, it was necessary to train additional data to enhance the annotation variations of each type of complex question.

This current experiment has showed that some candidates who did not have the highest probability value were selected by the system to become the best candidate. These candidates still had an Expected Answer Type (EAT) and question focus. Accordingly, these incorrect decomposed questions need to be tested on the baseline Indonesian Factoid QAS. This experiment was intended to know that these questions were still able to provide the correct results. For future development, the system also should accommodate complex question which has multiple conjunctions.

6. References

[1]. Arai, K., A. Handayani, N. (2012): Question Answering System for Effective Collaborative Learning, *International Journal of Advance Computer Science and Application (IJACSA)*.

[2]. Chali, Y., dan Hasan, S. A. (2012): Simple or Complex? Classifiying the Question by the Answer Complexity, *Workshop on Question Answering for Complex Domain*.

[3]. Corinna, C., dan Vapnik, V.. (1995): Support-Vector Network, *AT&T Bell Labs – USA*.

[4]. Harabagiu, S., Lacatusu, F., Hickl, A. (2006): Answering Complex Question with Random Walk Model, *SIGIR Conference on Research and Development in Information Retrieval*.

[5]. Kalyanpur, A., S., Patwardhan, B. K., Boguraev, Lally A., dan Chu-Carroll, J. (2012): Fact-based Question Decomposition in DeepQA, *IBM Journal on Research and Development*.

[6]. Kilicoglu, H., Fiszman, M., dan D.D Fushman. (2013): Interpreting Consumer Health Question, *Proceeding of BioNLP Wokshop*.

[7]. Lacatusu, F., Hickl, A., dan Harabagiu, S. (2006): Impact of Question Decomposition on the Quality of Answer Summaries, *Proceeding of LREC*.

[8]. Mitchell, T. (1997): *Machine Learning*, McGraw-Hill Series in Computer Science.

[9]. Purwarianti, A., Tsuchiya, M., Nakagawa, S. (2007): A Machine Learning Approach for Indonesian Question Answering System, *Proceeding of the IASTED International Conference on Artificial Intelligence and Its Application (AIA 2007)*, 6 pages, 12-14 February 2007, Innsbruck, Austria

[10]. Purwarianti, A., Tsuchiya, M. and Nakagawa, S. (2007): A Machine Learning Approach for an Indonesian-English Cross Language Question Answering System, *Journal of IEICE Transactions on Information and Systems*, pp. 1841-1852, Volume E90-D No 11, November 2007

[11]. Roberta, K. V., Wisudawati, L. M., Razi, M., dan Agushinta, D. (2011): Web Based Virtual Agent for Tourism in Indonesia, *Proceeding of Advance Computing and Communication (ACC)*.

[12]. Roberts, K., Masterson, K., Fiszman, M., Kilicoglu, H., Demner-Fushman, D. (2014): Annotating Question Decomposition on Complex Medical Question, *Proceeding of Language Resource and Evaluation (LREC)*.

[13]. Roberts, K., Fiszman, M., Kilicoglu, H., Demner-Fushman, D. (2014): Decomposing Consumer Health Question, *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP)*.

[14]. Zulen, A. A, dan Purwarianti, A. (2012): Study and Implementation of Monolingual Approach on Indonesian Question Answering System for Factoid and Non-Factoid Question, *Proceeding of CYBERNETICSCOM*.

[15]. Zulen, A. A, dan Purwarianti, A. (2013): Using Phrase Based Approach in Machine Learning Based Factoid Indonesian Question Answering, *Proceeding of CISAK*.

He received Sarjana (Bachelor) in Informatics Engineering from Telkom Institute of Technology (ITTelkom), Indonesia, in 2007. He received Master Degree in Informatics Engineering from Bandung Institute of Technology (ITB), Indonesia 2015. His research interests include the Natural Language Processing and Voice Processing. He is currently working as lecturer in Informatics Department University of Muhammadiyah Malang (UMM).

She was graduate from her bachelor and master degree at Informatics/Computer Science Program, Bandung Institute of Technology. She got her doctoral degree from Toyohashi University of Technology, Japan. Since 2008, she has become a lecturer at School of Electrical Enginering and Informatics, Bandung Institute of Technology, Indonesia. Her research interest is on computational linguistics, mainly on Indonesian natural language processing. She is now active as education officer at IEEE Indonesia and she is also active at Indonesian Association for Computational Linguistics.